What's in a looking time? [Short summary of main conclusion]

Christina Bergmann[1], Hugh Rabagliati[2], Sho Tsuji[3,4]

1: Max Planck Institute for Psycholinguistics

2: The University of Edinburgh

3: Ecole Normale Superieure, EHESS, CNRS, PSL University

4: The University of Tokyo

Author note

This research was supported by grants from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 659553 to the last author, and the Agence Nationale pour la Recherche [ANR-17-EURE-0017].

Corresponding authors:

Sho Tsuji, International Research Center for Neurointelligence, The University of Tokyo, 7-3-1 Hongo Bunkyo-ku, Tokyo, 113-0033 Japan, Contact: tsujish@gmail.com

Christina Bergmann, Max Planck Institute for Psycholinguistics, Wundtlaan 1, 6525 XD Nijmegen, Netherlands, Contact: chbergma@gmail.com

Hugh Rabagliati, The University of Edinburgh, 7 George Square, Edinburgh EH8 9JZ, United Kingdom, Contact: hugh.rabagliati@ed.ac.uk

Data availability statement:

All study materials, raw data, protocols, and scripts are available on the project website on the Open Science Framework at https://osf.io/f2v53/?view_only=df520a418e764552af528174905ac7b3.

[DOI to be added when the repository is made public after anonymized peer review is complete]

Research Highlights

- Infant looking time paradigms can yield mixed patterns of novelty and familiarity preferences.
- This paper aims to clarify whether both directions of preference contain comparable amounts of evidence.
- Using meta meta-analytic methods, we assess the relative evidential value of more versus less-frequent preferences.

Abstract

Looking time preference methods are a ubiquitous tool for tapping into infants' early skills and knowledge. However, predicting what preference infants will show in these paradigms can be difficult, and studies investigating the same ability oftentimes report opposing patterns of preference. For example, most studies investigating infant pattern learning report preferences for novel stimuli, but some report preference for familiar stimuli. How should such differences in preference direction be interpreted? One possibility is that any statistically significant preference is evidence for discrimination, such that all preferences provide similar evidential value. But another possibility is that the less-frequent preferences are so-called "sign errors", a form of sampling error in which a result is statistically significant, but the estimated effect size has the incorrect sign, e.g., showing a familiarity rather than novelty preference. In this paper, we use meta-analytic methods and statistical modeling to examine whether, when literatures show a heterogeneous pattern of looking time preferences, those preferences provide consistent evidential value, or whether one direction of preference may be a sign error.

[Summary of included meta-analyses, results, implications]

Keywords: Looking time, preference, sign error, meta-analysis

What's in a looking time? [Short summary of main conclusion]

The scientific study of infants' cognitive development has offered unique insights into the nature of early knowledge, and the mechanisms that drive learning and developmental change. Since the middle of the last century, methods relying on looking times or orientation direction (which we refer to collectively as looking time methods) have become the dominant experimental paradigm for providing these insights, and have been used to test questions on a variety of topics, from visual and auditory perception, to linguistic, conceptual, and social development.

Oftentimes, the measure of interest is infants' relative *preference* for one type of stimulus over another, and variations on this method have been used to establish central findings in the infant cognition literature, including infants' preference for infant-directed over adult-directed speech (Fernald, 1985), their ability to segment words from fluent speech (Jusczyk & Aslin, 1995), their ability to learn abstract patterns (e.g., Marcus, Vijayan, Rao, & Vishton, 1999), and their preference for faces of the same race (Kelly et al., 2005). The ubiquity of these methods is due to a variety of reasons. First and foremost, they can be used with very young infants, enabling researchers to experimentally study cognition in participants who cannot yet crawl, speak, or walk. Orientation responses, such as in head-turn procedures (e.g., Fernald, 1985) can be elicited in infants as young as 4 months (see Johnson & Zamuner, 2010), and gaze responses can be obtained even from newborns (e.g., Turati & Simion, 2002). Preference methods have also become popular because they are low-cost, requiring little more than a few video cameras, speakers, and computers. Finally, preference methods are versatile, allowing researchers to study a wide variety of questions using many different stimuli; for instance, simple preference tests can detect infants pre-existing biases

towards particular stimuli (e.g., infant- over adult-directed speech; Cooper & Aslin, 1990), and these can be augmented with an exposure phase in order to study how infants learn about particular stimuli or react to stimulus changes.

However, one potential concern about these methods is that their key dependent measure – the direction of a preference – can be hard to interpret, particularly when the paradigms used are quite complex. The interpretation of classic work using these methods (e.g., form perception, Fantz, 1961; Nelson, et al., 1995) was comparatively simple, suggesting that infants will preferentially orient toward images or sounds that are more distinct or interesting, and that they will learn to maintain their responses when the continuation of stimulation is contingent on their behavior. However, when more complex experimental designs are used, the interpretation of infant preferences has proved more difficult.

For example, in important work on how infants segment words from fluent speech, Jusczyk and Aslin (1995) familiarized infants to two words spoken in isolation, and then assessed if they preferred to listen to sentences that contained these now-familiar words, or control sentences that, instead, contained two different words. In a design like this, it is not clear which of the two types of test sentences should have been more interesting for the participants; in this study, infants preferred the sentences containing familiar words, a finding that has frequently been replicated (see Bergmann & Cristia, 2016). Conversely, other studies have found consistent preferences for more novel stimuli. In a study on how infants learn abstract patterns, Marcus et al. (1999) familiarized infants to sound sequences following a particular sequential rule, e.g., *pa-pa-di* and *wo-fe-fe* follow an ABB rule. They then assessed whether infants preferred to listen to new sequences in which novel sounds followed the familiar pattern (e.g., *gu-gu-ta*) or in which novel sounds followed a novel pattern (e.g.,

gu-ta-gu). Based on Jusczyk and Aslin (1995), one might have expected infants to prefer the former stimuli, which mixed novel and familiar elements, but in fact they preferred the latter, a result that has, again, been frequently replicated within the same context (see Rabagliati, Ferguson, & Lew-Williams, 2019). Thus, these two sets of findings illustrate how challenging it is to predict the preference infants might show in a looking time paradigm.

One might expect that it would be simpler to predict the direction of preference in studies that use what is called a habituation design, in which infants are first exposed to a stimulus up to a criterion of inattention, and then tested on whether they recover their attention when confronted with novel stimuli (for a review of classic work on infant visual habituation, see Colombo & Mitchell, 2009, and for an example of applying this method to study language learning, see Stager & Werker, 1997). By and large, the expected response is thus a novelty preference. However, for some studies that use habituation designs, it can still be difficult to generate predictions because, in these studies, all test trials are somewhat novel. In Fiser and Aslin's (2002) study of visual statistical learning, infants were habituated to co-occurring triplets of shapes, but then on test trials showed a preference for pairs of shapes that had been associated during the habituation phase, rather than pairs that had not been associated. That is to say, for these novel test stimuli (shape pairs rather than triplets), infants attended longer to the items that were still somewhat familiar. Thus, even in habituation designs, predictions about the direction of infants' preferences can be difficult to make, and it is yet unclear what might cause familiarity preferences in habituation studies.

The difficulty of predicting and interpreting these differences in directionality would be little more than a curio if it were the case that infants uniformly show particular preferences for particular paradigms, e.g., always prefer familiarity in word segmentation paradigms, and always prefer novelty in pattern learning paradigms. However, this is

empirically not the case. In the word segmentation literature, the preference for familiar stimuli predominates, but some studies still report longer orientation times for sentences containing novel words (e.g., Singh, 2008). Meanwhile, although most follow-ups to Marcus et al.'s (1999) study of pattern learning reported a novelty preference, studies such as Johnson et al. (2009) found a significant preference for familiar patterns. Thus, in well-matched tasks, infants sometimes show a direction of preference opposite from the majority of experiments, and it is as yet unclear why such shifts in preference direction occur. The goal of the present study is to understand how we should treat these findings. Specifically, we will evaluate two main possibilities that have been put forward in the literature. On the one hand, it has been suggested that these flips in preference within a literature could represent meaningful data about infant cognition that can inform theories of development. For example, changes in direction of preference might indicate that infants respond differently as they mature, or when confronted with stimuli of higher or lower complexity. Consequently, such results would deserve further experimental and theoretical consideration. On the other hand, it is also possible that (some of) these changes are instead artifacts, driven by a combination of factors such as publication bias, researcher degrees of freedom, and the noisy measurements typically used in infant studies. Under this scenario, researchers should carefully consider how to treat new evidence that is not in line with a majority direction of preference, discounting it as positive evidence while still placing it in the public record, to allow for unbiased accumulation of evidence.

The first possibility, that flips in preference direction are driven by theoretically meaningful factors, is an exciting one, since it implies that the results of preference methods could give us quite fine-grained information about how infants process stimuli. Perhaps the most relevant theoretical model here is Hunter and Ames (1988), who proposed that

familiarity, novelty, and null preferences can be explained by the assumption that infants who are exposed to a given stimulus will continue to explore it until they have completely encoded it. In particular, infants who have a choice between a partially encoded stimulus and a completely novel stimulus will prefer the more familiar one. However, once infants have fully encoded this familiar stimulus, they are prepared to start encoding novel information, switching to preferring the more novel stimulus. These assumptions lead to two predictions for experimental outcomes. First, with other factors constant, prolonging familiarization time should eventually lead to a novelty preference. Second, lower task complexity should lead to an earlier switch from familiarity to novelty preference. These predictions also relate to age, such that, for a given task, older infants will switch to a novelty preference earlier compared to younger infants, due to their more mature cognitive capacities and more extensive experiences with complex input (see also Roder et al., 2000; Houston-Price & Nakai, 2004; Sirois & Mareschal, 2002).

Models such as Hunter and Ames (1988) appear to well-characterise how infants respond in visual habituation tasks, which assess how infants' attention to a simple visual stimulus declines over time, and how that attention recovers when a different simple visual stimulus is substituted (Colombo & Mitchell, 2009; Oakes, 2010). The evidence is less consistent for studies that use more complex stimuli, or that use habituation to assess cognitive processes other than visual attention and memory. A recent meta-analysis of infant word segmentation (Bergmann & Cristia, 2016) did not find evidence that novelty versus familiarity preferences could be predicted by factors such as cognitive load or age. In particular, their analysis of 168 experiments from 51 articles and with data gathered from 3774 infants, did not find any evidence that a switch can be predicted based on age, native language, or various task- and stimulus-related factors that might affect difficulty. The

authors concluded that the Hunter and Ames model could not explain why some studies in the word segmentation literature occasionally report a novelty preference.

Nevertheless, there are a variety of meaningful methodological factors that could systematically affect infants' preference for novel or familiar stimuli. Importantly, these factors may often be hard to detect. Infant behavior in looking time experiments has been suggested to be affected by a variety of subtle, often unreported, methodological factors, such as lighting or noise (Johnson & Zamuner, 2010), which could impact how well a given stimulus is processed, and thus whether a particular stimulus will generate a familiarity or novelty preference. This could cause systematic differences in behavioral responses to similar types of stimuli. For example, laboratories that test their participants in soundproofed rooms might be less susceptible to noise and therefore give infants the opportunity to process stimuli differently, consequently resulting in a different pattern of preferences than those laboratories whose testing rooms do not shield against ambient noise.

This line of reasoning has the important consequence that all preferences, regardless of directionality, should be interpretable, as they would represent evidence that infants can discriminate between stimuli (for more in-depth discussion see e.g., Aslin, 2007; Houston-Price & Nakai, 2004; Oakes, 2010). That is to say, if infants in a quiet lab provide evidence for learning patterns through a novelty preference, but infants in a more noisy lab provide evidence for learning patterns through a familiarity preference, then in both cases infants have still provided evidence for learning. As Burnham and Dodd (1998) note "From one point of view, it can be argued that the valence of the preference is unimportant, for as long as there is a preference, then discrimination is demonstrated" (p.174). Thus, under this proposal, all types of infant preference should be similarly informative about the phenomenon being investigated, even if we might not be able to directly trace back the source

of a particular preference direction in practice. And indeed, previous studies on word segmentation or rule learning that reported unexpected preference directions (e.g., Johnson et al., 2009; Singh, 2008) tended to interpret those preferences as providing robust evidence for discrimination and learning.

Alternatively to the position held above, namely that both directions of preference are informative and meaningful, it is also possible that, for any given paradigm, only one direction of preference is informative about the phenomenon under test. Recent statistical considerations suggest that perhaps not all preferences are created equal, and that so-called non-dominant preferences may provide notably less, or no, evidential value compared to the dominant, typically-reported preference. In particular, non-dominant preferences may be a so-called sign error (Gelman & Carlin, 2014), which is to say, a statistically significant result in the opposite direction of the true underlying effect. Such sign errors are statistically inevitable, and become more likely in cases of low statistical power, which is often true for infant research (Bergmann et al., 2018). Publication bias would then lead to **o**nly sign errors associated with significant *p* values becoming part of the public record. By contrast, null results, particularly in the non-standard direction of preference, which would all stem from a single underlying effect, would not be published due to the same publication pressures. Therefore, in combination, the relative absence of null results, which remain in the filedrawer, and the presence of significant findings in two directions might create an impression that both directions of preference are inherently meaningful.

In the framework provided by Gelman and Carlin (2014), sign errors do not occur with a very high frequency, except under conditions of extremely low power (when they are similarly frequent to true positive results); thus, the fact that switches in preference are (at least impressionistically) quite frequent in the infant cognition literature might be taken to

suggest that sign errors cannot be the only explanation for changes in preference. However, two factors militate against this conclusion. The first factor is the finding that studies of infant cognition often have very low power (Bergmann et al., 2018), which would naturally lead to higher rates of sign errors. The second factor is researcher degrees of freedom (e.g., Simmons, Nelson, & Simonsohn, 2011). In particular, we would suggest that sign errors in practice are likely more prevalent than in Gelman and Carlin's theoretical framework, because that framework assumes that unbiased analytic decisions. However, there is ample evidence (e.g., Sterling, 1959) that statistical analyses in the published literature are likely to have a number of biases, such as a bias to publish studies that show a statistically significant outcome, which will increase the proportion of exaggerated (i.e. statistically significant) results in either direction.

Consistent with this reasoning, recent work suggests that non-dominant preferences may not be indicators of meaningful evidence. In a recent meta-analysis of the infant rule learning literature, Rabagliati, Ferguson, and Lew-Williams (2019) used a technique called *p*-curve analysis (Simonsohn, Nelson, & Simmons, 2014; 2015) to examine the evidence provided by non-dominant preferences. *P*-curves assess the evidential value in a literature by analyzing how statistically significant *p* values are distributed across a set of studies on the same research question: In literatures that are strongly contaminated by publication bias such that only Type 1 errors are published, this distribution should be uniform between 0 and .05 (because the distribution of *p* values under the null hypothesis is uniform). Rabagliati and colleagues found that, for studies reporting a non-dominant preference for familiarity, the distribution of *p* values did not differ from the uniform distribution, suggesting that the reported results reflected publication bias rather than a true effect. Moreover, the evidential value of these results (operationalized as their estimated statistical power) was significantly

less than the evidential value from those studies reporting the dominant preference for novelty. These findings thus provide some initial evidence that, at least within this specific literature, different directions of preferences provide different amounts of evidence.

In the present study, we will provide a more representative and generalizable test of whether or not differences in preference directions within different infant literatures provide meaningful evidence or are statistical artefacts that were strengthened by publication pressures. We substantially extend the preliminary findings of Rabagliati and colleagues (2018) to a far broader selection of the literature on infant cognition, conducting *p*-curve analyses on existing meta-analyses, in order to provide a clear test of whether preferences in the non-dominant direction provide evidential value, or whether these non-dominant preferences are more likely to represent sign errors. We use two types of *p*-curve analyses, described in more detail in the Methods section. First, we conduct separate *p*-curve analyses on each individual meta-analysis, similarly to the method used in Rabagliati, Ferguson, and Lew-Williams (2019), which allows us to assess whether, for each topic represented by a meta-analysis, the dominant direction of preference provides greater evidential value than the non-dominant direction. Second, we conduct a multi-level analysis, allowing us to aggregate evidence across different meta-analyses, to test whether it is generally true that the dominant direction of preference provides greater evidential value than the non-dominant direction.

Our analyses will speak to key aspects of theory building, assessing the intrinsic meaningfulness of changes in the direction of preference, and whether researchers are best-advised to interpret such changes as important evidence, or as sampling artifacts. Furthermore, in the case that our analyses find that changes in direction of preference are indeed meaningful, our study would invite new experimental and meta-analytic investigations, to determine which precise mechanism might cause such a change. Our results

will also inform current statistical practice, providing insights as to whether more high-powered one-sided tests in the presence of a majority direction of preference are permissible, and whether meta-analyses should rely on positive versus negative or absolute effect sizes (see Cristia, 2018).

Methods

Additional materials are available on the project website on the Open Science Framework at https://osf.io/f2v53/?view_only=df520a418e764552af528174905ac7b3.

Sample Characteristics

Our unit of observation is individual research studies that we aim to gather by collating all available systematic meta-analyses on infant looking or listening preference. Thus, in this paper, meta-analysis refers to a collection of experimental studies that has been compiled by clear selection rules (although note that we might analyze a subset of those studies included in a published meta-analysis, see our additional selection criteria below). Relying on extant meta-analyses allows us to build on previous work in assembling and coding comprehensive sets of studies on a literature of interest.

To gather all available meta-analyses, we will first extract all suitable meta-analyses from the MetaLab database (metalab.stanford.edu; Bergmann et al., 2018), and will conduct two literature searches using Google Scholar and PsychINFO: one for the terms "meta-analysis [title]" + "infant" + "visual preference", and one for the terms "meta-analysis [title]" + "infant" + "listening preference". To illustrate the potential size of our evidence base, the first search yielded 288 results on Google Scholar on January 15th, 2019, the second

31. In addition, all authors of this paper added meta-analyses they had become aware of prior to the time of writing in an expert list, which currently includes 6 entries. The project page will contain a complete search protocol and full screening spreadsheet; after removing duplicates, all meta-analyses will be screened for inclusion (see Figure 1 and subsection Inclusion criteria below).

Table 1 will list all included meta-analyses and their key characteristics, including citation, topic, source (MetaLab, literature searches, expert list), infant age (mean and range in months), and the number of effect sizes in each direction of preference (not filtered by statistical significance and only those that are statistically significant in parentheses). To ensure reproducibility, we will share both the data as well as our pre-selection scripts, additional codings of all meta-analyses, search protocols, and further materials on the project website.

Inclusion criteria

Suitable meta-analyses will be those in which the effect sizes for the reported studies contrast preferences for two kinds of stimuli. For example, this could include time spent orienting toward one type of stimulus versus another, as in the Head-turn Preference Procedure, or the time spent fixating one type of stimulus over another, as in many modern eye tracking paradigms. Some meta-analyses combine various methods, for example neurocognitive and behavioral measures. In these cases, we will only use the subset of the meta-analysis where procedures measuring behavioral preference have been used. We will further screen all studies within the included meta-analyses to avoid studies appearing twice in the analyses pooling over multiple meta-analyses (for example because it tests an interaction between two

phenomena examined in two different meta-analyses). As a rule, we will remove the study from the larger meta-analysis.

For a specific meta-analysis to be included, it must contain at least six significant effects from peer-reviewed studies in either direction. This cut-off is based on statistical considerations, since we require two effects for each random parameter of the partial pooling analysis (see Planned Analyses section). We focus on significant effects because publication bias is a key issue in our study. There is no recommendation of a minimal number of significant effects for conducting *p*-curves, but we are confident that the minimum of six *p* values will yield reasonable estimates in our analyses.

The last columns of Table 1 will indicate the number of (significant) outcomes per direction of preference for each meta-analysis. Figure 1 will show how we arrived at our final set of meta-analyses included.

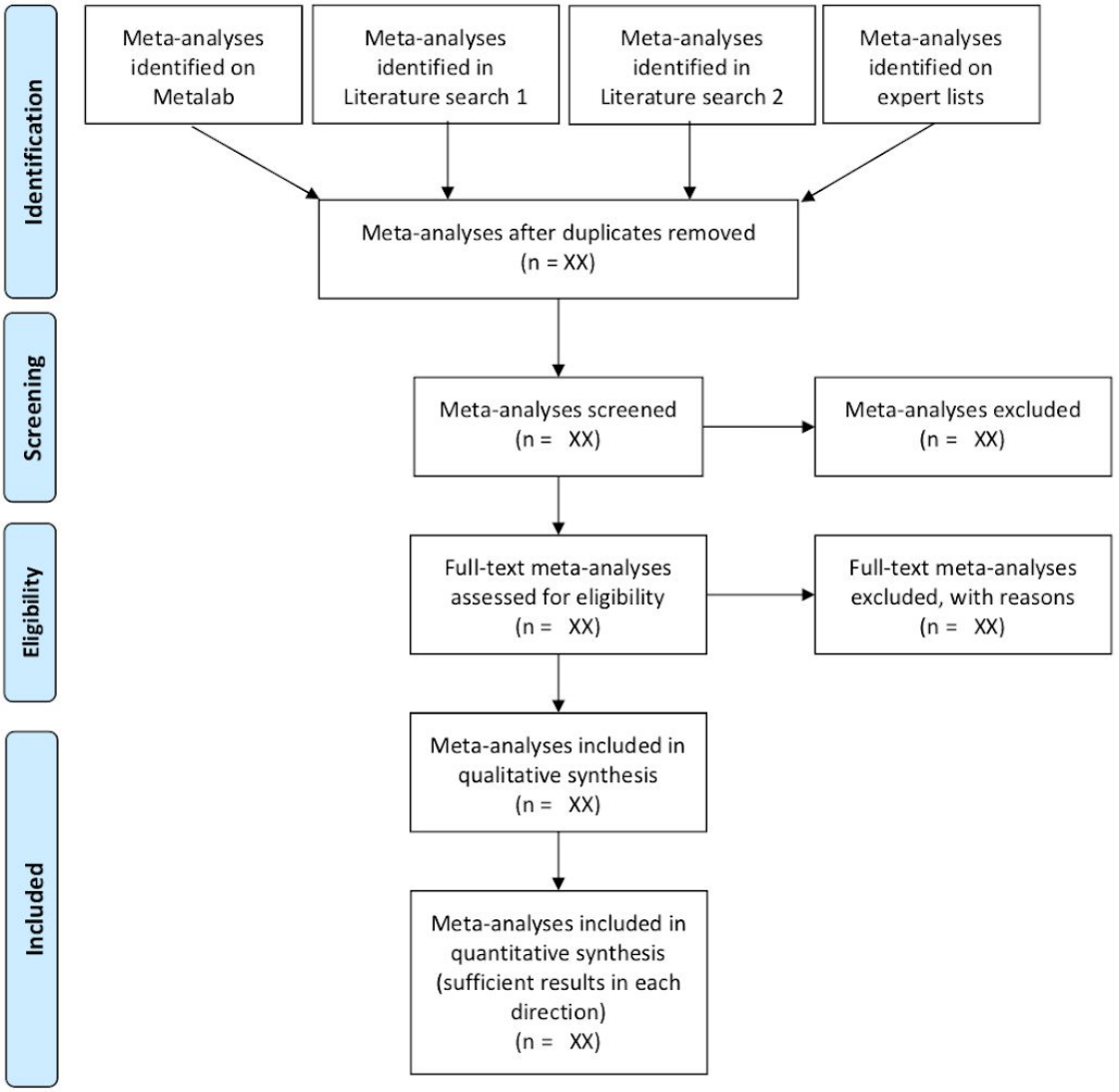

Figure 1. Flow-chart of the meta-analysis search and exclusion pipeline.

In our analyses, we will not exclude any data as "outliers". While outlier exclusion is important for unbiasing standard meta-analytic methods, we suggest that it is not useful in this case, as our goal is to evaluate the strength of the published evidence, and so if a *p*-value is published, it necessarily counts as such evidence.

Table 1. Overview of the included meta-analyses.

Citation	Topic	N Effect	Infant age	N effect	N effect	Source

		sizes	in months (range)	sizes familiarity (significant)	sizes novelty (significant)	
Authors (year)	Topic 1	100	8 (6-12)	50 (20)	50 (10)	MetaLab
Authors (year)	Topic 2	70	24 (16-36)	20 (3)	50 (25)	Literature search 1

Procedure

Data collection and annotation. Our study selection rule includes all systematic meta-analyses that we could identify through searching google scholar and PsychINFO, through the MetaLab database of developmental meta-analyses, and through an expert list, which use a behavioral preference procedure that relies either on head-turns or looking times.

All meta-analyses in this study will be annotated in a common coding scheme to make meta-meta-analyses possible. In particular, we will code for each included experiment: the details of the method (e.g., head turn preference), test population characteristics (participant number, infant age), and test statistics (raw means and standard deviations, t-values, F-scores). We will also report an effect size: Since the relevant experimental procedures result in a preference score, that effect size will be drawn from the family of standardized mean differences, and we intend to use Hedges' *g*, which is a de-biased variant of Cohen's *d* and which is preferable when sample sizes are small (Hedges, 1981).

For the *p*-curve analyses, we will further document specific characteristics of each included experiment in a *p*-curve disclosure table based on information from the source paper to ensure that the studies analyzed are comparable (Simonsohn, Nelson, & Simmons, 2014). This table contains a study identifier, a quote of the original hypothesis from the source

paper, the study design, the single key statistical result to test the previously mentioned hypothesis, and a quote from the source paper including this statistical result. Contra standard recommendations, we will not include results of robustness tests because these are not commonly reported in the infant literature. We add a number of columns to accommodate the specific research question: source; experiment condition (to be able to include multiple effect sizes per paper when appropriate); the expected outcome (significant or not); the expected direction (familiar, novel, or none; either based on a statement in the paper or inferred from the seminal paper); the actual outcome (significant or not, coded to accommodate reporting errors; see Nuijten et al., 2016); the observed direction of preference (familiar, novel, or none); a quote from the source paper interpreting the observed outcome and direction of preference; and whether there was a change in the direction of preference across a series of studies within a paper (if the paper contained multiple studies). We also code the reported *p* value. We further note whether studies gathered from previous meta-analyses were peer-reviewed, and limit our analyses to peer-reviewed studies only. There are two main reasons to limit our analyses to peer-reviewed studies. First, on the practical side, not all meta-analyses include studies that were not peer reviewed (for example as a means of quality control), and it is not always transparently reported whether efforts were made to uncover unpublished data at all (Polanin, Hennessy, & Tsuji, 2019). Second, on the conceptual side, we propose that publication bias would be the underlying cause of non-dominant directions of preference being sign errors, rather than genuine effects (and thus we would expect only very small numbers of unpublished, significant non-dominant preferences).

Each study, extracted from a source paper, will be coded by one of the authors of this paper. For reliability checking, we will independently recode a subset of 5 studies per meta-analysis. Because of the size of our overall dataset, we will provide the full tables in the

supplementary materials and on the project website. Table 2 illustrates how we plan to code studies for an example meta-analyses.

Table 2. Excerpt from the *p*-curve disclosure table. See text for a detailed description.

Citation	Exp	Cond	Description	Predicted Outcome	Predicted Direction	Change in Preference	Reported *p* value
Author1 (year)	E1	A	Infants should discriminate F from N	Significant	None	None	0.017
Author2 (year)	E1	A	Young infants are predicted to prefer F over N.	Significant	Familiarity	None	0.04
Author2 (year)	E1	B	Older infants are predicted to prefer N over F.	Significant	Novelty	Fam_to_ nov	0.03
Author2 (year)	E2	NA	Infants in control condition will show no preference.	Non-significant	None	None	> .05

Planned analyses

Our analyses rely on the *p*-curve meta-analytic technique developed by Simonsohn, Nelson and Simmons (2014, 2015). As described in the Introduction, *p*-curve analyses are used to evaluate the amount of evidence that is present within a literature, which they do by examining the distribution of statistically significant results in that literature. When the null hypothesis is true, *p* values are uniformly distributed between 0 and 1 (hence, there is a one in twenty chance of a Type 1 error when applying an alpha threshold of .05). But when a literature systematically tests a null hypothesis that is false, then distribution of *p* values will

be right skewed, with more values close to 0. The degree of right skew will be larger for literatures that use higher powered tests, for example because they are analyzing effects that are larger or because the measures are more precise. Thus, the key logic behind *p*-curve analyses is to estimate the degree to which *p* values in a literature deviate from a uniform distribution, and thus provide evidential value. In this way, *p*-curve offer a complement to standard meta-analytic techniques such as meta-regression. Whereas those techniques aim to estimate the size of a typical effect in a literature, and assess which factors might moderate the size of that typical effect, *p*-curve analyses aim to estimate whether a published literature provides evidential value, which allows us to ask a different question of importance. In the present case, *p*-curve allows us to compare studies that report opposite directions of effect sizes. For example, while a meta-regression analysis would be very likely to say that novelty and familiarity preferences are different from one another if we were to split studies by the direction of an effect, a *p*-curve analysis allows us to assess the degree to which they provide the same amount of evidence. The key question answered is thus not whether novelty and familiarity are different in their sign, but whether each set of results provides evidence of underlying effects in both directions.

Preprocessing: selecting and computing *p* values. *P*-curve analyses require quite stringent and specific data selection rules; for example, for any individual study, only one *p* value can be used (e.g., if a study reports a test of gaze behavior on the first pair of trials, and also on the second pair of trials, only one of these *p* values can be included). We will thus follow this guideline, based on the recommendations of Simonsohn, Nelson, and Simmons (2014): If multiple conditions are reported within an experiment of a paper, we will select the condition closest to the seminal experiment within a given meta-analysis (i.e. the condition

within a paper that did not introduce additional manipulations or used different stimuli). If multiple conditions are equivalent, we will use the first significance test reported.

Because *p* values are not always reported exactly, and because there might be errors in computing *p* values (see Nuijten et al., 2016; Hardwicke et al., 2018), we re-compute *p* values when possible from test statistics and degrees of freedom extracted from the source paper (i.e. *F* scores, *t*-values, etc). For cases where neither an exact *p* value nor test-statistics are available, we will attempt to re-compute test statistics and *p* values based on the reported effect sizes (which the authors of meta-analyses may have computed based on privately shared data) using the appropriate formulae. Note that for within-participant designs, deriving test statistics from standardized mean difference effect sizes (Cohen's *d* and Hedges' *g*) requires the correlation between the two measures, which is rarely reported. For each re-calculated *p* value, we will thus interpolate an intermediate correlation of .55 when it is not available.

Example scripts for our planned analyses can be found on the project's OSF repository.

Meta-analysis-specific analysis. In the meta-analysis-specific analysis, we will look at the studies included within each individual meta-analysis, and compare *p*-curves for studies that report the dominant versus non-dominant direction of preference. To give an example, based on the meta-analysis of word segmentation (Bergmann & Cristia, 2016) we will compare the *p*-curve of studies that report the more-frequent familiarity preference~~s~~ to the *p*-curve of studies that report the less-frequent novelty preference, while based on the meta-analysis of abstract rule learning (Rabagliati, Ferguson, & Lew-Williams, 2019) we will compare the *p*-curve of studies that report the more-frequent novelty preference to the

p-curve of studies that report the less-frequent familiarity preference. Thus, we will split each meta-analysis into two parts by direction of preference, conduct *p*-curve analyses on each part, and then compare the results of those analyses to determine whether one direction of preference contains greater evidential value.

We will conduct *p*-curve analyses using the procedure that can be found at p-curve.com, based on Simonsohn, Simmons, and Nelson (2015). This procedure is only performed on statistically significant results, and so meta-analyses are first trimmed of any reported findings that are not significant at the alpha level of $p < .05$; this is necessary because non-significant results are typically not reported, and so an analysis of all *p* values would be biased toward 0. Then, the probability of obtaining *p* values that are at least as extreme as each of the reported significant *p* values is calculated under two different scenarios. The first scenario, designed to test whether published findings contain significant evidential value, assesses the probability of obtaining *p* values at least as extreme as those reported if the null hypothesis were true (so that the expected distribution of significant *p* values is uniform between 0 and .05). These so-called *pp*-values (probability of *p* values) are then aggregated using a procedure known as Stouffer's method, which yields a *z* statistic, whose value will be greater when *p* values are more right skewed. A significance test can then be performed on the *z* statistic. The second scenario tests the hypothesis that the *p* values reflect a true effect, but that the sample size of the conducted experiments means that they have low power (33%). In this case, the expected distribution of *p* values is slightly right skewed, and we can test whether the resulting *p* values are less right skewed than this expected low-power distribution, in the same way as before.

We then assess if the evidential value from the non-dominant preference direction is smaller than the value from the dominant direction, by comparing the estimated statistical

power of each set of studies, again using a procedure from p-curve.com (Simonsohn, Nelson & Simmons, 2019). This procedure is similar to the procedure for calculating *pp*-values when power is estimated at 33% except that, instead, we now find the level of power that best fits the data by searching for the level that results in Stouffer's test yielding a *z* of 0 and *p* value equal to 0.5. Then, we compare power in the dominant set of studies to power in the non-dominant set, using bootstrapping to construct 95% confidence intervals around the difference in estimated power between the sets.

In sum, for each direction of preference within a meta-analysis, we will conduct a *p*-curve test of whether the data contains evidential value, and a *p*-curve test of whether the data contains significantly less than a small amount of evidential value. Then, we assess whether the dominant preference direction provides significantly more evidential value than the non-dominant direction. Note that these multiple analyses do not incorporate a correction for multiple comparisons, and so we urge readers to treat each individual statistical test with caution.

Partial pooling analysis. Our second analysis will aim to measure the distribution of *p* values found across our different meta-analyses, providing a global picture of the evidence provided by the dominant versus non-dominant direction of preference, and accounting for concerns about multiple comparisons. Importantly, the traditional method for conducting *p*-curve analyses (as described above) is not suitable for this purpose, because it does not obviously scale for use in a hierarchical fashion, such that we could assess sets of *p* values drawn from different types of experiments and asking different questions. Instead, we will use mixed effects regressions that model the *p* values across the studies in our meta-analyses as following Beta distributions, and then will assess whether the location of those

distributions differ between the more- versus less-frequent preference direction. Simplifying somewhat, the main aim of this mixed-effects analysis is to assess whether the typical significant *p* value for studies showing dominant preferences is closer to 0 than the typical significant *p* value for studies showing non-dominant preferences (cf., the suggestion in Simonsohn, Nelson & Simmons, 2015, p. 22); however, rather than use a linear mixed model, we use a Beta regression, as described below.

Beta regressions (Ferrari & Cribari-Neto, 2004; Figueroa-Zúñiga, Arellano-Valle, & Ferrari, 2013) are similar to generalized linear models such as logistic regression, except that the modeled response is assumed to be Beta distributed (see Figure 2A), with values falling within the standard unit interval (0,1), such that this type of regression can be used to model probabilities. This makes it more helpful for modeling *p* values, than other distributions. For example, a simple linear model would incorrectly assume that difference in evidence between a *p* value of .04 and .03, was greater than the difference in evidence between a *p* value of .01 and .00001, but that is not correct. Meanwhile, a logistic regression model would assume that *p* values can take the value 0 or 1, which they cannot. Finally, Beta distributions are flexible in the shapes that they can take: left-skewed, right skewed, uniform, and more, much as *p* values can have a variety of distributions, e.g., testing a true effect produces right skewed *p*s, testing a true null produces uniform *ps,* while extreme "*p*-hacking" can produce distributions that are left skewed or even bimodal, if *p*-hacking is done in the context of a true effect (cf. Simonsohn, Simmons, & Nelson, 2014). Beta regression can model all of these potential outcomes.

In Beta regression, the distribution of the response variable (here, *p* values) is predicted by two parameters, μ, which corresponds to the mean of the distribution, and ϕ, which is a precision parameter. Figure 2A shows how variation in the μ parameter can affect

the shape of the Beta distribution. These parameters can then be modeled as the linear combination of a set of predictor variables. In our case, the resulting parameters, particularly the μ parameter, should vary depending on differences in effect size, statistical power, and presence of publication bias/*p*-hacking. For example, when a set of *p* values comes from studies with high power to detect a true effect, so that the distribution is right skewed with many *p* values close to 0, then the μ parameter (i.e., the estimated mean of the distribution) should be low. When a set of *p* values comes from studies with lower power to detect an effect, such that *p* values are less right skewed, then the resulting μ parameter should be higher.

Proof of concept. Beta regressions are well-established in other fields (see the range of articles citing Ferrari & Cribari-Neto, 2004) but are rarely used in psychology, and to our knowledge have not been used to analyze *p*-values. We thus illustrate the method by reanalyzing the comparison between novelty and familiarity preferences in Rabagliati et al. (2018). There, the distribution of significant *p* values from studies reporting a novelty preference (the dominant direction of preference) was significantly more right skewed than the uniform, suggesting evidential value, while the distribution of *p* values from studies reporting a familiarity preference did not show a significant difference from the uniform (Figure 2B), providing no evidence of evidential value. In addition, the confidence intervals around the distributions from the two directions of preference did not overlap, suggesting that novelty preferences contained (more) evidential value compared to familiarity preferences.

To compare those two sets of significant *p* values, we first multiply them by 20 so that they are distributed within the interval (0,1), and then use a Beta regression of the form:

$$p \sim \text{Beta}(\mu,\phi)$$
$$\mu_i = \beta^{\mu}_{0} + \beta^{\mu}_{1} * \text{PreferenceType}_{[i]}$$
$$\phi_i = \beta^{\phi}_{0}$$

following the notation used by Gelman and Hill (2007). Here, *p* values follow a Beta distribution with location parameters μ and precision parameter ϕ. μ is modeled as a linear combination of predictors including an intercept and a predictor for PreferenceType that compares dominant novelty preferences to non-dominant familiarity preferences (with the latter set as the reference level). ϕ is modeled as an intercept. μ is fit using a logit link, and ϕ is fit using a log link. We implemented the model using the Stan statistical modeling platform (Carpenter et al., 2017; Sorensen, Hohenstein & Vasishth, 2016), via the R package BRMS (Bürkner, 2018). In BRMS syntax, the model has the form *bf(p ~ 1 + PreferenceType, phi ~ 1)*.

Figure 2C illustrates how the resulting model fits the data, and Figure 2D shows the results of 10 posterior predictive checks; i.e., simulations of the data generated from the model parameters. The model accurately reconstructs the right skewed distribution of novelty preference *p* values, and it estimates a uniform distribution for the familiarity preference *p* values, just as the original *p*-curve analysis did. Table 3 displays the three parameters of the model. The intercept of μ (i.e., the μ for non-dominant *p* values) is estimated at 0.08 with a wide 95% credible interval of -0.77 to 0.93 (consistent with the results reported in Rabagliati et al., 2018); recall that μ is fit with a logit link, so a value of 0 transforms to an estimated mean probability of 0.5, consistent with *p* values uniformly distributed between 0 and 1. Importantly, the change in μ for novelty *p* values is -1.41 [-2.38, -0.39], indicating that the mean is substantially lower, and thus that studies showing this preference provide stronger evidential value, again consistent with the results of Rabagliati et al. (2018).

Table 3: Results of a Beta regression on *p* values for novelty and familiarity preferences from Rabagliati et al. (2018)

Parameter	Estimate (Est. Error)	95% Credible Interval
μ Intercept (familiarity)	0.08 (0.43)	-0.77,0.93
μ PreferenceType (novelty)	-1.41 (0.51)	-2.38,-0.39
ϕ Intercept	0.45 (0.22)	0,0.87

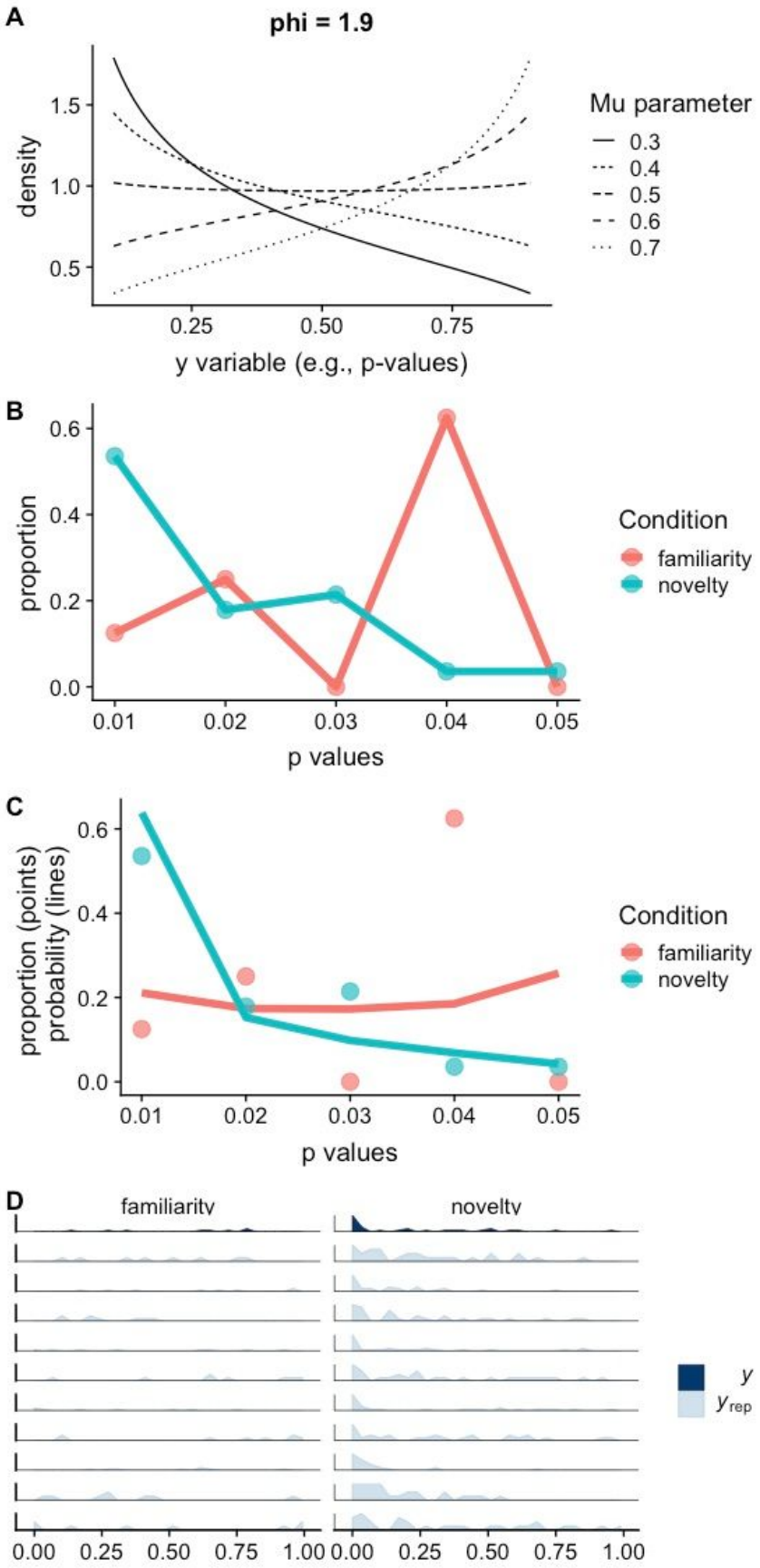
A
phi = 1.9
density
1.5
1.0
0.5
0.25
0.50
0.75
y variable (e.g., p-values)
Mu parameter
0.3
0.4
0.5
0.6
0.7
B
proportion
0.6
0.4
0.2
0.0
0.01
0.02
0.03
0.04
0.05
p values
Condition
familiarity
novelty
C
proportion (points)
probability (lines)
0.6
0.4
0.2
0.0
0.01
0.02
0.03
0.04
0.05
p values
Condition
familiarity
novelty
D
familiarity
novelty
0.00
0.25
0.50
0.75
1.00
y
y rep

Figure 2. Graphs illustrating the Beta regression analysis of *p* values. A. An illustration of possible shapes that the Beta distribution can take. B. *P*-curves for novelty and familiarity preferences in studies investigating infant abstract rule learning. C. The estimated distribution of *p* values via Beta regression. D. Posterior predictive checks illustrating the fit of the Beta regression to the distribution of *p* values.

Planned analyses. Our first analysis will compare evidential value across the more- versus less-frequent preference directions. We plan to model our overall dataset with a hierarchical Beta regression of the form

$$p \sim \text{Beta}(\mu,\phi)$$

$$\mu_i = \beta^{\mu}_{0} + u_{[\text{meta-analysis[i]}]} + \beta^{\mu}_{1} * \text{PreferenceType}_{[i]} + u_{1[\text{meta-analysis[i]}]}$$

$$\phi_i = \beta^{\phi}_{0} + u^{\phi}_{[\text{meta-analysis[i]}]}$$

Compared to the example Beta regression above, there are three new parameters: μ is further predicted by a random intercept for each meta-analysis (u) and a random slope for the effect of preference type for each meta-analysis (u_1), and ϕ is predicted by a random intercept u^{ϕ}. In BRMS syntax, the model has the form *bf(p ~ 1 + PreferenceType + (1|meta-analysis), phi ~ 1 + (1|meta-analysis))*.

If the planned model does not converge, we will first attempt to fit it by varying the sampling parameters (e.g., increasing the number of iterations or increasing the target acceptance rate by changing the adapt_delta parameter). If this is not successful, we will refit the model using priors that are more conservative; for example, the default priors for population-level effects in models constructed using BRMS are flat over the real numbers, and we would replace these with normal priors, centered on 0, with a standard deviation of 5.

Follow-up analyses

Subsequent analyses will examine how methodological and study design features might moderate difference in evidential value. These follow-up analyses will necessarily have lower power than our main analysis, and so we necessarily have to treat them speculatively.

First, we will analyse whether the effect of preference direction on evidential value varies depending on the type of exposure method used. We will compare habituation tasks, where familiarization terminates when an individual infant reaches a predetermined habituation criterion, to tasks with a fixed familiarization period, and tasks with no pre-exposure. The key hypothesis is that unexpected preference directions may contain more evidential value for methods other than habituation. Habituation methods are specifically designed to elicit novelty preferences, and so non-dominant preferences (i.e., familiarity) are more likely to be caused by sampling error than by experimental design features. By contrast, with other methods, it is harder to predict the direction of preference, and thus, at a baseline level, both types of preference may contain evidential value. This analysis will follow the same form as the planned Beta regression, but the fixed effect of preference direction will interact with a further set of fixed effect predictors (dummy coded) for exposure method.

Subsequent analyses will test whether the evidential value for non-dominant preferences increases when we account for theoretically motivated factors. We will test whether non-dominant preferences contain more evidential value when authors state that they predicted the preference direction ahead of time, based on the coding described in Table 2. In particular, we assess whether evidential value is higher when the author stated that they predicted the non-dominant preference. This analysis will again follow the same form as our first planned Beta regression, but a fixed effect of prediction statement will interact with the fixed effect of preference direction.

Finally, we will assess two factors that, in theory, could predict non-dominant preferences, according to well-known models of infant preference direction such as Hunter and Ames (1988). First, we will analyse whether the effect of preference type interacts with participant age (centered per meta-analysis), based on the assumption that non-dominant preferences may occur more likely outside of the standard age range of a paradigm (e.g., if a paradigm typically analyses 12-month-old infants, and the typical result is a novelty preference, then familiarity preferences might be more robust when shown by 8-month-olds than by 10-month-olds).

Second, we will analyse effects of stimulus complexity. If dynamic and complex stimuli, due to increased task complexity, indeed are more likely to lead to mixtures of preferences than static stimuli, that will engender different preferences depending on age. For each record, we will code whether the test stimulus is a static visual image or a more complex stimulus (e.g., animations, linguistic stimuli). We will incorporate this fixed effect predictor in our Beta regression, testing whether the effect of preference type interacts with this dummy-coded predictor for stimulus complexity.

Consideration of statistical power. Our analysis plan relies on a combination of frequentist and Bayesian inference. Standard *p*-curve analyses rely on hypothesis tests, while interpretation of our Beta regressions depends on inspection of credible intervals; sample size plays an important role in both of these assessments. However, for this project, consideration of sample size is somewhat moot because our analyses will include all currently available and suitable meta-analyses of paradigms that assess infant preferences. Increasing the number of available meta-analyses would require us to conduct new meta-analyses, which simply is not feasible.

Importantly, prior work suggests that we should have a large enough sample size to generate reasonable statistical inferences. Rabagliati et al.'s work on infant preferences during rule learning showed that evidence from novelty preferences could be distinguished from evidence from familiarity preferences in a sample consisting of 54 significant **studies** in total. We estimate that our sample will be substantially times larger than this, providing us with a far clearer picture.

Positive controls. To ensure that our method distinguishes whether shifts in preference reflect meaningful data or statistical artifacts, two tests are necessary. First, we need to ensure that our method reliably detects evidential value given our sample sizes, and in particular that our method can detect evidential value in the smaller samples that will arise for non-dominant preferences. To assess this, we will conduct a resampling analysis, in which we draw samples from the dominant preference data that are matched in size to the samples from the non-dominant preference data, and test if we can still detect evidential value in samples of this smaller size. Thus, this test will assess whether our results are a consequence of numerical differences in sample size.

Second, we want to assess whether our sample size allows us to detect the degree of publication bias and so-called p-hacking (which leads to left skewed *p* values; Simmons, Nelson & Simonsohn, 2011). To do this, we will conduct additional simulation studies (similar to those reported in the supplementary materials) in which we simulate data drawn from distributions that are either uniform (no evidential value) or left-skewed (p-hacked to obtain "just significant" *p* values), and assess whether our sample size allows us to reliably draw conclusions about the degree of skew. The results of these positive controls will be

reported in detail in the supplementary materials, and we will summarize them briefly in the main paper.

Completed Work at Time of Submission for Stage 1 Peer Review

In order to get a grasp of the workload involved in this project we coded a selection of meta-analyses prior to submission. To this end, we first created a coding template listing required details from each meta-analysis for our analyses, basing ourselves on recommendations in Simonsohn, Nelson, and Simmons (2014). The template is available on our project page.

We entered the data from three meta-analyses on MetaLab into this template, with two more in progress. We chose these meta-analyses randomly from those that fulfill the criterion of testing behavioral preference over one over another type of stimulus in the visual or auditory domain. One of these is the rule learning meta-analysis (Rabagliati et al., 2018), which has previously been p-curved, but not undergone our second beta regression analysis. The other were meta-analyses of statistical sound category learning (Cristia, 2018), phonotactic learning (Cristia, 2018), infant sound symbolism (Fort, Lammertink, Peperkamp, Guevara-Rukoz, Fikkert, & Tsuji, 2018), and word segmentation (Bergmann & Cristia, 2016).

For each of these meta-analyses, we coded or are in the process of coding each row that provided a p value for the relevant comparison of preference to one over the other type of stimulus. We estimate that coding each of these rows takes an average of 5 minutes. Based on an average meta-analysis size of 84 rows in MetaLab, we therefore estimate that coding of one meta-analysis would take 420 minutes or 7 hours.

Timeline for study completion

We estimate a maximum of 12 months for literature search, coding, and reliability checks (dependent on the number of meta-analyses we can include), 2 months for analysis, and another 2 months for completing the results. We will thus aim to submit the stage 2 report 16 months after acceptance of the stage 1 report.

References

Aslin, R.N. (2007). What's in a look? *Developmental Science, 10* (1), 48–53. DOI: 10.1111/j.1467-7687.2007.00563.x

Bergmann, C., & Cristia, A. (2016). Development of infants' segmentation of words from native speech: A meta‐analytic approach. *Developmental Science, 19*(6), 901-917. DOI: 10.1111/desc.12341

Bergmann, C., Tsuji, S., Piccinini, P. E., Lewis, M. L., Braginsky, M., Frank, M. C., & Cristia, A. (2018). Promoting replicability in developmental research through meta‐analyses: Insights from language acquisition research. *Child Development*, *89*(6), 1996-2009. DOI: 10.1111/cdev.13079

Bürkner, P.-C. (2018). Advanced Bayesian Multilevel Modeling with the R Package brms. *The R Journal, 10*(1), 395-411.

Burnham, D., & Dodd, B. (1998). Familiarity and novelty in infant cross-language studies: factors, problems, and a possible solution. *Advances in Infancy Research*, *12*, 170-187.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software 76*(1). DOI: 10.18637/jss.v076.i01

Colombo, J., & Mitchell, D. W. (2009). Infant visual habituation. *Neurobiology of learning and memory*, 92(2), 225-234.

Cooper, R. P. & Aslin, N. R. (1990). Preference for Infant-directed Speech in the First Month after Birth. *Child Development. 61*(5), 1584-1595. DOI: 10.1111/j.1467-8624.1990.tb02885.x

Cristia, A. (2018). Can infants learn phonology in the lab? A meta-analytic answer. *Cognition*, *170*, 312-327. DOI: 10.1016/j.cognition.2017.09.016

Fantz, R. L. (1961). The origin of form perception. *Scientific American*.

Fernald, A. (1985). Four-month-old infants prefer to listen to motherese. *Infant Behavior and Development*, *8*(2), 181-195. DOI: 10.1016/S0163-6383(85)80005-9

Ferrari, S., & Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics, 31*(7), 799-815. DOI: 10.1080/0266476042000214501

Figueroa-Zúñiga, J. I., Arellano-Valle, R. B., & Ferrari, S. L. (2013). Mixed beta regression: A Bayesian perspective. *Computational Statistics & Data Analysis, 61,* 137-147. DOI: 10.1016/j.csda.2012.12.002

Fiser, J., & Aslin, R. N. (2002). Statistical learning of new visual feature combinations by infants. *Proceedings of the National Academy of Sciences, 99*(24), 15822-15826. DOI: 10.1073/pnas.232472899

Fort, M., Lammertink, I., Peperkamp, S., Guevara‐Rukoz, A., Fikkert, P., & Tsuji, S. (2018). Symbouki: a meta‐analysis on the emergence of sound symbolism in early language acquisition. *Developmental Science*, *21*(5), e12659. DOI: 10.1111/desc.12659

Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science*, *9*(6), 641-651. DOI: 10.1177/1745691614551642

Hardwicke, T. E., Mathur, M. B., MacDonald, K., Nilsonne, G., Banks, G. C., Kidwell, M. C., ... & Lenne, R. L. (2018). Data availability, reusability, and analytic reproducibility: Evaluating the impact of a mandatory open data policy at the journal Cognition. *Royal Society open science*, *5*(8), 180448. DOI: 10.1098/rsos.180448

Hedges, L.V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational and Behavioral Statistics, 6*, 107–128. DOI: 10.3102/10769986006002107

Houston-Price, C., & Nakai, S. (2004). Distinguishing novelty and familiarity effects in infant preference procedures. *Infant and Child Development, 13*(4), 341–348. DOI: 10.1002/icd.364

Hunter, M.A., & Ames, E.W. (1988). A multifactor model of infant preferences for novel and familiar stimuli. *Advances in Infancy Research, 5*, 69–95.

Johnson, S. P., Fernandes, K. J., Frank, M. C., Kirkham, N., Marcus, G., Rabagliati, H., & Slemmer, J. A. (2009). Abstract rule learning for visual sequences in 8-and 11-month-olds. *Infancy*, *14*(1), 2-18. DOI: 10.1080/15250000802569611

Johnson, E., & Zamuner, T. (2010). Using infant and toddler testing methods in language acquisition research. In E. Blom & S. Unsworth (Eds.), Language learning & language teaching: Vol. 27. *Experimental methods in language acquisition research* (pp. 73-93). Amsterdam, Netherlands: John Benjamins Publishing Company. DOI: 10.1075/lllt.27.06joh

Jusczyk, P. W., & Aslin, R. N. (1995). Infants′ detection of the sound patterns of words in fluent speech. *Cognitive Psychology*, *29*(1), 1-23. DOI: 10.1006/cogp.1995.1010

Kelly, D. J., Quinn, P. C., Slater, A. M., Lee, K., Gibson, A., Smith, M., ... & Pascalis, O. (2005). Three‐month‐olds, but not newborns, prefer own‐race faces. *Developmental Science, 8*(6), F31-F36.DOI: 10.1111/j.1467-7687.2005.0434a.x

Marcus, G. F., Vijayan, S., Rao, S. B., & Vishton, P. M. (1999). Rule learning by seven-month-old infants. *Science*, *283*(5398), 77-80. DOI: 10.1126/science.283.5398.77

Nelson, D. G. K., Jusczyk, P. W., Mandel, D. R., Myers, J., Turk, A., & Gerken, L. (1995). The head-turn preference procedure for testing auditory perception. *Infant Behavior and Development, 18*(1), 111-116. DOI: 10.1016/0163-6383(95)90012-8

Nuijten, M. B., Hartgerink, C. H. J., van Assen, M. A. L. M., Epskamp, S., & Wicherts, J. M. (2016). The prevalence of statistical reporting errors in psychology (1985-2013). *Behavior Research Methods*, *48 (4)*, 1205-1226. DOI: 10.3758/s13428-015-0664-2

Oakes, L.M. (2010). Using Habituation of Looking Time to Assess Mental Processes in Infancy. *Journal of Cognition and Development, 11*(3), 255-268, DOI: 10.1080/15248371003699977

Polanin, J., Hennessy, E., & Tsuji, S. (2019). Transparency and reproducibility of meta-analyses in psychology: A meta-review. *Manuscript under review.*

Rabagliati, H., Ferguson, B., & Lew‐Williams, C. (2019). The profile of abstract rule learning in infancy: Meta‐analytic and experimental evidence. *Developmental Science, 22*(1), e12704. DOI: 10.1111/desc.12704

Roder, B.J., Bushnell, E.W., & Sasseville, A.M. (2000). Infants' preferences for familiarity and novelty during the course of visual processing. *Infancy, 1*(4), 491–507. DOI: 10.1207/S15327078IN0104_9

Segal, J. & Newman, R.S. (2015) Infant Preferences for Structural and Prosodic Properties of Infant-Directed Speech in the Second Year of Life. *Infancy. 20*(3), 339-351. DOI: 10.1111/infa.12077

Shufaniya, A., & Arnon, I. (2018). Statistical Learning Is Not Age‐Invariant During Childhood: Performance Improves With Age Across Modality. *Cognitive Science, 42*(8), 3100-3115. DOI: 10.1111/cogs.12692

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*(11), 1359-1366.

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General, 143*(2), 534-547. DOI: 10.1037/a0033242

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2019). p-curve app (4.06). Retrieved from http://www.p-curve.com/app4/

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2015). Better P-curves: Making P-curve analysis more robust to errors, fraud, and ambitious P-hacking, a Reply to Ulrich and Miller (2015). *Journal of Experimental Psychology: General, 144*(6), 1146-1152. DOI: 10.1037/xge0000104

Singh, L. (2008). Influences of high and low variability on infant word recognition. *Cognition, 106*(2), 833–870. doi: 10.1016/j.cognition.2007.05.002

Sirois, S., & Mareschal, D. (2002). Models of habituation in infancy. *Trends in Cognitive Sciences, 6*(7), 293–298. DOI: 10.1016/S1364-6613(02)01926-5

Stager, C. L., & Werker, J. F. (1997). Infants listen for more phonetic detail in speech perception than in word-learning tasks. *Nature, 388*(6640), 381-382. DOI: 10.1038/41102

Sterling, T. D. 1959. Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa. *Journal of the American Statistical Association, 54*(285), 30–34. DOI: 10.1080/01621459.1959.10501497

Turati, C., & Simion, F. (2002). Newborns' recognition of changing and unchanging aspects of schematic faces. *Journal of Experimental Child Psychology*, *83*(4), 239-261. Doi: 10.1016/S0022-0965(02)00148-0

Sorensen,T., Hohenstein, S., & Vasishth, S. (2016). Bayesian linear mixed models using Stan: A tutorial for psychologists, linguists, and cognitive scientists. *Quantitative Methods for Psychology*, 12(3), 175-200. DOI: 10.20982/tqmp.12.3.p175

List of Figure Legends

Figure 1. Flow-chart of the search and exclusion pipeline.

Figure 2. Graphs illustrating the Beta regression analysis of *p* values. A. An illustration of possible shapes that the Beta distribution can take. B. *P*-curves for novelty and familiarity preferences in studies investigating infant abstract rule learning. C. The estimated distribution of *p* values via Beta regression. D. Posterior predictive checks illustrating the fit of the Beta regression to the distribution of *p* values.

List of Table Legends

Table 1. Overview of the included meta-analyses.

Table 2. Excerpt from the *p*-curve disclosure table. See text for a detailed description.

Table 3. Results of a Beta regression on *p* values for novelty and familiarity preferences from Rabagliati et al. (2018)